Russian verbal aspect and machine translation

Zangenfeind, Robert ; Sonnenhauser, Barbara

ВИД РУССКОГО ГЛАГОЛА И МАШИННЫЙ ПЕРЕВОД

Цангенфайнд Р.И. (r.zangenfeind@lmu.de), Зонненхаузер Б. (basonne@lmu.de)
Мюнхенский университет, Мюнхен, Германия

Ключевые слова: вид глагола, машинный перевод, неоднозначность, семантический/синтаксический признак, русский язык, английский язык, немецкий язык, турецкий язык

RUSSIAN VERBAL ASPECT AND MACHINE TRANSLATION

Zangenfeind R. (r.zangenfeind@lmu.de), Sonnenhauser B. (basonne@lmu.de)
University of Munich, Munich, Germany

Abstract

Rule-based machine translation still offers some very beneficial facets for linguistic theory, because by implementing rules on the computer linguistic theory can be verified in practice. One of the most intricate problems for machine translation is grammatical aspect in Russian when it has to be translated into a language either lacking aspect or having a different aspect system. On the categorical level, aspect has only approximate equivalents in non-Slavic languages, such as the progressive form in English, for instance. In addition, language-internally, its semantics and interpretation cannot be sufficiently captured with only one specific characteristic feature. In this paper, we aim at establishing a basis for the machine translation of the Russian aspect. To do so, we discuss an approach to describe the interaction of verb and aspect semantics in a systematic way. Moreover, we describe a possible annotation for the aspectual information that is provided by further lexical components contributing to the meaning computation. This allows for the formulation of rules for machine translation into target languages where the grammatical category of aspect is realized differently or not present at all.

Keywords: grammatical aspect, machine translation, ambiguity, semantic/syntactic features, Russian, English, German, Turkish

0. Introduction

While statistical machine translation has made great progress over the last years, rule-based machine translation still offers some very beneficial facets. The great virtue of formulating and implementing rules for machine translation instead of using a pure statistical approach is that a rule-based approach is a precious source for theoretical linguistics, cf. Iomdin (2003, 2008), and Apresjan et al. (1989:285).

If dictionaries and rules are implemented in an appropriate way, the computer will be able to produce correct translations. If it does not, it is obvious that the dictionaries or rules have to be improved or new rules have to be added to the system. Thus, the knowledge of rules that describe natural language will be widened and the theory of linguistics will be augmented. This means that even "negative linguistic material" in the form of incorrect translations of a rule-based machine translating system will help to improve linguistic theory.

Apresjan et al. (1989:285) point out that the computer makes mistakes of a different kind from those that a human translator makes. Thus, unique negative linguistic material is provided. Iomdin (2003) gives an example how erroneous automatic parsing of a Russian sentence leads to a wrong translation into English. The examination of this sentence and its syntactic structure reveals a special syntactic property of a group of Russian nouns (*ideja* 'idea' etc.), concerning copulative sentences, that another group of nouns (*tsel'* 'purpose' etc.) doesn't have. By introducing a specific syntactic feature for the according lexemes the parser can be fixed and the sentence is translated correctly.

An especially difficult problem for machine translation is the analysis of the various meanings of Russian verbal aspects. This is a field where rule-based machine translation can be very helpful if appropriate rules are formulated, implemented and verified at the computer. In this paper, we want to discuss the problems of language-internal aspect interpretation and present steps towards rules for machine translation of aspect, especially from Russian to English.

1. Rules for aspect?

Since aspect interpretation is context-driven and to a large degree subject to pragmatic reasoning, a statistical approach runs into troubles from the very beginning. Gaining statistically valid results for all the possible interpretations would require an immensely large parallel corpus. This makes a rule-based approach look more promising. However, formulating rules for the interpretation and translation of Russian aspect is a rather intricate problem for at least two reasons: this is a highly polysemous category, as can be seen from the

numerous readings and sub-readings listed in grammars and textbooks for both aspects, and it has hardly one-to-one correspondences in other aspect languages.

1.1 Interpretation and correspondences

The multiple interpretations for the imperfective (ipf) aspect can be classified, among others, as 'actual-processual', 'conative', 'habitual', 'atemporal', 'general-factual' and 'durative'. Some readings for the perfective (pf) aspect are the 'event' reading, the 'perfect' and the 'pluperfect' reading. These readings are largely influenced by context. But even considering context, it is not always clear, which interpretation to choose, i.e. which interpretation might be the 'right' one. This makes it quite hard to formulate a common semantic basis for the pf and ipf aspect.

Grammatical aspect is present in other languages as well, e.g. in English and Turkish: At first sight, English *-ing* and Turkish *-iyordu* (*-iyor*=progressive, *du*=past) seem to correspond to the ipf aspect, which would leave the English simple form and the Turkish unmarked past (*-di*) as equivalents to the pf aspect. Such correspondences would simplify the problem of machine translation a lot. But while English uses the progressive form for the actual-processual reading, there is no one-to-one correspondence in the other cases. The habitual interpretation is rendered by the simple form, cf. (1), as is the durative reading, cf. (2). The same holds for the atemporal and the general-factual interpretation, while the conative reading can only be expressed by lexical means.

(1) *At night he **played** with guitarist Luther Perkins and bassist Marshal Grant.* (http://en.wikipedia.org/wiki/Johnny_Cash, 9.1.2014)

(2) *From 1969 to 1971, Cash **starred** in his own television show […]* (http://en.wikipedia.org/wiki/Johnny_Cash, 9.1.2014)

Pretty much the same holds for Turkish: *-iyordu* is used for the actual-processual interpretations, the unmarked past for the durative and general-factual interpretations. In addition, Turkish has one further aspect marker, which is used for atemporal and habitual readings, the so-called 'aorist', cf. (3):

(3) *Daha 4 sene öncesine kadar Play Station'da sırf Gerrard'ı kontrol etmek için Liverpool'u **seçerdim**, şimdi beraber oynuyorum.* (Luis Suarez; http://fotogaleri.hurriyet.com.tr, 10.1.2014) 'Until four years ago I chose Liverpool on

the Play Station, just to have Gerrard under control, now we play together.'

As regards the Russian pf aspect, it is expressed in English and Turkish mainly in terms of tense.

Thus, even though English and Turkish have a morphological category of aspect, there is no one-to-one correspondence to Russian. Comparing the semantic range of the Russian, English and Turkish aspect markers, we get the relations illustrated in table 1. German, which does not have a morphological category of aspect, has to rely on lexical and syntactic means:

<table>
<tr><th>Russian</th><th>Turkish</th><th>English</th><th>German</th></tr>
<tr><td>pf</td><td rowspan="2">-di</td><td rowspan="2">simple form</td><td rowspan="3">Ø</td></tr>
<tr><td rowspan="2">ipf</td></tr>
<tr><td>-ir(di)</td></tr>
<tr><td></td><td>iyor(du)</td><td>-ing</td><td></td></tr>
</table>

Table 1: Relations of aspect markers in different languages

In order to be able to eventually formulate rules in an 'if-then'-format, thus, the following two main problems have to be solved: (i) specify the 'if'-part by language-internally figuring out the relevant interpretation, and (ii) specify the 'then'-part by cross-linguistically figuring out the corresponding equivalent expression. The prerequisite for both is a well-formulated semantic description of aspect.

2. Aspect semantics

Since it is not possible for to rely on formal equivalences, translation has to take into account the content side. What machine translation cannot achieve is the transfer of specific interpretations since these take into account also extra-linguistic knowledge. What machine translation can achieve, is the transfer of semantically coded meanings. This amounts to the difference between polysemy as the availability of various interpretations for one form and ambiguity as the existence of clearly distinct meanings for one and the same formal expression. This is well-known also from lexical semantics.[1] What is needed in a first step is, thus, a semantic analysis of aspect that is able to distinguish between ambiguity and

[1] Cf. the German form *Bank* which has at least three meanings: 'bank', 'bench' and 'river bank'. Each of these meanings has its own range of interpretations, i.e. 'bank' may be interpreted as the financial institution, the building, the system, and the like. When it comes to translation, it is not these specific interpretations that are crucial but the three distinct meanings.

polysemy.

2.1 Polysemy and ambiguity

One possible way of systematizing aspect interpretations in terms of ambiguity and polysemy is provided by the analysis developed in Sonnenhauser (2004, 2006), based on the combination of a selection-theoretic (Bickel 1996) and time-relational (Klein 1995) account. According to this analysis, aspect operators select, and thereby assert, specific part(s) of the event structure encoded by the verb. Assuming a tripartite event-structure (Moens, Steedman 1988), verbs may encode (i) dynamic phases 'φ_{dyn}' (preparatory processes), (ii) boundaries 'τ' (culmination points) and (iii) static phases 'φ_{stat}' (consequent states), depending on the eventuality they refer to. By selecting and asserting some part of the coded event structure, aspect establishes a relation between the topic time interval I(TT) (the time the assertion is about) and the event time interval I(e) (that part of the run time of the denoted event that is selected by the aspect operator).

The pf aspect can be described by the fact that the boundaries of the event-structure are included in the topic time (a more detailed account is provided in Sonnenhauser 2006, 2009). These boundaries are specified in the course of interpretation: the interval may be closed to both sides, i.e. the initial and final points are part of the interval, it may be open to the right or open to the left, i.e. the initial point is part of the interval whereas the final point is excluded and vice versa. This is illustrated with the example in (4a), which can be interpreted in three ways and thus be translated into English as in (4b–d):

(4) a. *Ja emu **dala** knigu.*

b. *I **gave** him the book [and then …]* I(TT) closed

c. *I **have given** him the book [and now …]* I(TT) open to the right

d. *[After] I **had given** him the book* I(TT) open to the left

For the ipf aspect the following relations between topic time interval and event time interval are relevant:

(5) a. $I(TT) \subset I(\varphi_{dyn})$

*Kogda on voshel, ona **chitala** knigu.* 'When he came in, she **was reading** a book.'

($I(\varphi_{dyn})$: the time interval of her reading the book, covering only this process

excluding beginning or end; I(TT) is included in the reading-process and specified by the moment when he came in)

b. I(TT) = I(e)

*Ona **rabotala** v universitete.* 'She **worked** at the university.' [= She was employed there.]

(I(e): the time interval when she was employed at the university; I(TT) runs exactly parallel to the time interval of her working at the university)

c. I(TT) ⊃ I(e)

*Ona uzhe **rasskazyvala** emu ètu istoriju.* 'She **has** already **told** him this story.'

(I(e): the time interval of her telling the story; I(TT) includes the complete story-telling event)

It is these ambiguities that are decisive for the purposes of machine translation; both the structures underlying the representations and the specific interpretations can be neglected.

2.2 Cross-linguistic evidence

The justification for postulating the three specifications for the pf aspect is provided not only on language-internal grounds, but also by the fact that these relations can be morphologically coded in other languages, which render it mainly in terms of temporal distinctions. Table 2 illustrates this for Russian, English and German, with the brackets indicating the boundedness-characteristics of the intervals. Note that these correlations hold for the past tense.

semantics	**interpretation**	**Russian**	**English**	**German**
group I_{pf} TT closed: [---τ---]	eventive	pf	simple past	imperfect / perfect
group II_{pf} TT right open: [---τ---[	perfect (existential, current relevance, extended now, etc.)	pf	perfect	perfect
group III_{pf} TT left open:]--- τ---]	pluperfect	pf	pluperfect	pluperfect

Table 2: Ambiguity of pf aspect

Likewise, the cross-linguistic validity of assuming three basic ipf configurations is suggested by two facts: the three configurations may be coded morphologically in other languages in terms of aspect distinctions, and if coded, they give rise to a similar range of interpretations. This is illustrated in table 3, comparing 'imperfective' grammemes in Russian, English and

Turkish (for more details cf. Sonnenhauser 2006).[2] This indicates that even though aspect is grammaticalized in all three languages, they are by no means equivalent as regards the semantic range of the respective grammemes.

semantics	**interpretation**	**Russian**	**English**	**Turkish**
group I_{ipf} TT $\subset \varphi_{dyn}$	processual, conative	ipf	progressive	-iyordu -mekteydi
group II_{ipf} TT = e	habitual, non-actual, potential, permanent, atemporal	ipf	simple form	-irdi
group III_{ipf} TT $\supset$ e	general-factive, durative	ipf	simple form	-di

Table 3: Ambiguity of ipf aspect

The ambiguity of the Russian aspects and the cross-linguistic validity of the possible disambiguated configurations are crucial for the question of machine translation in that this provides the basis for stating clearly formulated rules.

2.3 Disambiguation

Disambiguation is achieved by specifying I(TT) in terms of its boundedness-features and – for the ipf aspect – by specifying the relevant part of the Aktionsart that is selected and related to this interval. In Russian, this specification is possible mainly by lexical and syntactic means: as regards the ipf aspect, adverbs like *medlenno* 'slowly' or *postepenno* 'gradually' specify I(TT) as open-bounded, adverbs like *ran'she* 'formerly' as unbounded, particles like *uzhe* 'already' as closed-bounded, and hence the interpretation as belonging to group I_{ipf}, II_{ipf}, or III_{ipf} respectively. Concerning the pf aspect, conjunctions like *i* 'and [then]' disambiguate eventive (group I_{pf}) from perfect (group II_{pf}) interpretations, cf. (6a) vs. (6b), adverbials specifying a point in time suggest the pluperfect interpretation (group III_{pf}), cf. (6c), etc.:

(6) a. *Ja **otkryl** mashinu i sel*. (NKRJa) 'I **opened** the car and [then] got in.'
[---τ---]

b. *Zato synok eë v gorode magazin **otkryl**. Vot i radujtes'* (NKRJa)
'Instead, her son **has opened** a shop in the city. So be glad…'
[---τ---[

c. *On uzhe **otkryl** rot, no tut v komnatu shirokim shagom voshel djadja Kolja.*

[2] The comparison in table 3 is confined to the past, since group III_{ipf} is not possible for the other tenses. Accordingly, the Turkish forms are specified with the past tense morpheme *-di*.

(NKRJa) ‘He already **had opened** the mouth, but there uncle Kolja entered the room with big steps.’

]---τ---]

As can be seen from tables 2 and 3, for machine translation from Russian to English, German or Turkish it is enough to solve these basic ambiguities. What is rendered by means of the perfect in English or German has the same interpretational range as the ‘perfect’ / group II_{pf} specification of the Russian pf aspect, what is rendered by means of the *-irdi* suffix in Turkish may give rise to the same variety of interpretations as group II_{ipf} of the Russian ipf aspect. The same reasoning applies to the other ambiguities.

For an automatic disambiguation, the relevant lexical and syntactic means have to be annotated in the lexical entries of lexemes as regards the aspectual information they contribute to the meaning computation. The computation may then proceed in the form of ‘if-then’ statements along the lines proposed by Vazov (1999), which is also used by Mel’chuk, Wanner (2008) for aspect-establishing rules in the process of German-Russian translation.

3. Towards rules for aspect

The machine translation system ÈTAP-3[3] makes use of a system of semantic and syntactic features (e.g. ‘DLIT’ to characterize a period of time) which provide a lot of information for lexemes that can be useful for the interpretation of aspect.

For our purpose this system of features could be enriched with a part of the classification of predicates by Apresjan (2006). This classification includes 17 classes. Some of them exclude certain disambiguation possibilities and/or make others highly probable. For ‘dejatel’nosti’ (‘activities’)[4], such as *torgovat’* ‘to trade’, for instance, the actual-processual and the general-factual readings are ruled out, whereas a durative interpretation is most likely. Other classes, such as ‘dejstvija’ (‘actions’), are a lot less explicit and allow for all possible interpretations. For their disambiguation, further information provided by other aspectually relevant components in the regarded sentence must be taken into account.

Adverbials, particles and conjunctions provide this information.[5] These parts of speech have

---

[3] ÈTAP-3 is a rule-based MT system for translations from Russian to English and vice versa, and also includes some further NLP applications (cf. Apresjan et al. 2003).

[4] The English terms for classes of predicates are taken from Apresjan (2005).

[5] These components correspond to the contextual clues (imperfective and perfective triggers) of Mel’chuk, Wanner (2008).

to be assigned with additional semantic and syntactic features respectively in their lexical entries. Another crucial bit of information is provided by tense. Present tense, for instance, excludes ipf interpretations out of group III_{ipf} and all pf interpretations except for the future interpretation. The combination of all this kind of information can be the basis for the "calculation" of a temporal and aspectual interpretation of the whole sentence.
An example to illustrate which information in a sentence is relevant is given in (7):

(7) *Ran'she ja po vecheram prodelyval èti gimnasticheskie uprazhnenja po pjat' raz.*[6]
lit. 'formerly I in evenings do.PAST.ipf these gymnastic exercises each five times'

Most lexemes and phrases in this sentence are important for our interpretation. For all of them the dictionary entries of ÈTAP already provide some important information, which, for our purposes, should be enriched by the following:

- *ran'she* 'formerly' is temporally and referentially (as concerns reference to event) indefinite and thus excludes group I_{ipf} interpretations; appropriate semantic features could be 'temporally indefinite' and 'referentially indefinite'[7]

- *po [vecheram]* 'in [the evenings]': the preposition in this expression – governing a temporal lexeme in the dative case, i.e *po16*[8] – expresses regularity. An adverbial phrase like *po vecheram* 'in the evenings' can be annotated by labeling the preposition *po16* with the feature 'regularity'; thus, it excludes group I_{ipf} and group III_{ipf} interpretations

- *prodelyvat'* '[to] do' is used as a support verb; i.e. it has no semantics, only its aspectual information (=ipf) is relevant

---

6 Example from Bendixen et al. (2005–2012).

7 The semantic feature 'temporally indefinite' indicates that there is just a vague temporal specification in terms of localization on the time axis. The lists of adverbs with this and other semantic features still must be thoroughly examined; the need for a list of such triggers is pointed out also by Mel'chuk, Wanner (2008:141). 'Referentially indefinite' concerns the selection and assertion of a specific part of the event structure carried out by aspect (cf. section 2.1): adverbs like *ran'she* indicate that there is no specific part of the event structure selected by aspect (some more examples of such features are given in Sonnenhauser, Zangenfeind 2013).

8 cf. Slovar' russkogo jazyka (1983).

- *uprazhnenie* ‘exercise’ is the semantic predicate in the sentence and can be labeled as ‘zanjatie’ (‘occupation’) according to Apresjan (2006: 83, 86f.); in combination with an ipf support verb such as *prodelyvat’* it allows for group I_{ipf}, II_{ipf} and III_{ipf} interpretations

- *po [pjat’ raz]* ‘[five times] each’: the preposition here – governing a noun that can have a numeral as syntactic dependent, i.e. *po20*[9] – expresses distributivity of the verbal complement and allows for group I_{ipf}, II_{ipf}, III_{ipf} interpretations; the preposition *po20* can be labeled with the feature ‘distributive’.

Based on this information, the aspectual information given in (7) can be calculated and, thus, disambiguated as follows:

(8) for language-internal disambiguation:
IF predicate has feature ‘occupation’
AND IF aspect = ipf
AND IF tense = past
AND IF there is an adverb of ‘group II_{ipf}’
THEN ‘group II_{ipf}’ interpretation

(9) for translation into English:
IF ‘group II_{ipf}’ interpretation
THEN ‘simple form’ in English[10]

Formal descriptions like these can be the basis for an implementation in a machine translation system like ÈTAP.[11]

---

9 cf. Slovar’ russkogo jazyka (1983).

10 The most adequate translation would be with the habitual construction ‘used to’. This specification can be solved by means of language-internal paraphrasing rules and is not necessarily an immediate concern of translation.

11 Since ÈTAP includes a highly developed Russian-to-English MT system, we intend to implement rules for aspect translation into English in a first step. But beginnings for the implementation of Russian-German translation in ÈTAP have already been made and are developed further by R. Zangenfeind and others. So, in the long run the translation of aspect from Russian to German is also planned.

4. Conclusion

In machine translation a rule-based approach for the interpretation and translation of the Russian verbal aspect looks promising when using the combination of a selection-theoretic and time-relational account to systematize the semantics of aspect and its interpretations. This systematization comprises several groups specifying the relation between topic time interval and event time interval. Disambiguation of the semantics of aspect is made possible by annotating all relevant lexemes with specific, aspectually relevant information. This is the starting point for a possible computational implementation of aspect interpretation. Enriching the system of semantic and syntactic features of the machine translation system ÈTAP with Apresjan's classification of predicates and with additional, more detailed syntactic/semantic features, we discussed the problems of a "calculation" of aspect interpretation and presented steps towards a possible solution.

Our future work will be to develop the necessary system of semantic features for verbs and predicative nouns, adverbials, particles and conjunctions. It is our aim to implement rules for aspect translation in the machine translation system ÈTAP. Besides the practical utility, an implementation in a rule based system has the great virtue to verify the linguistic theory in practice and, with that, to enable an improvement of the theory.

References

Apresjan Ju.D. (2005), Prolegomena to systematic lexicography, in: Apresjan Ju.D., Iomdin L.L. (eds.), East West encounter: second international conference on Meaning ↔ Text Theory, Moscow, pp. 20–29.

Apresjan Ju.D. (2006), Fundamental classification of predicates [Fundamental'naja klassifikatsija predikatov], in: Apresjan, Ju.D. (ed.), A linguistic picture of the world and systematic lexicography [Jazykovaja kartina mira i sistemnaja leksikografija], Moscow, pp. 75–110.

Apresjan Ju.D. et al. (1989), Linguistic Software for ÈTAP-2 System [Lingvisticheskoe obespechenie sistemi ÈTAP-2], Moscow.

Apresjan Ju.D. et al. (2003), ÈTAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the Meaning <=> Text Theory, in: Conference Proceedings of MTT 2003, Paris, pp. 279–288, available at http://proling.iitp.ru/publications/.

Bendixen B. et al. (2005–2012), Russian up to date [Russisch aktuell], Wiesbaden.

Bickel B. (1996), Aspect, mood and time in Belhare, Zürich.

Dictionary of the Russian Language (1983), vol. III [Slovar' russkogo jazyka, Tom III], Moscow.

Iomdin L. (2003), Purpose and Idea: a Lesson Drawn from Machine Translation, in: Conference Proceedings of MTT 2003, Paris, pp. 269–278.

Iomdin L. (2008), A few Lessons Learned from rule-based Machine Translation, in: Gross G., Schulz K.U. (eds.), Linguistics, Computer Science and Language Processing. Festschrift for Franz Guenthner on the Occasion of his 60th Birthday. London, pp. 171–187.

Klein W. (1995), A time-relational analysis of Russian aspect, in: Language 71(4), pp. 669–695.

Mel'chuk I.A., Wanner L. (2008), Morphological Mismatches in Machine Translation, in: Machine Translation, 22, pp. 101–152.

Moens M., Steedman M. (1988), Temporal ontology and temporal reference, in: Computational Linguistics 14(2), pp. 15–28.

Sonnenhauser B. (2004), Underspecification of 'meaning': the case of Russian imperfective aspect, in: Proceedings of the ACL-04 workshop on 'Text Meaning and Interpretation'. Barcelona, pp. 89–96.

Sonnenhauser B. (2006), Yet there's method in it. Semantics, pragmatics, and the interpretation of the Russian imperfective aspect, Munich.

Sonnenhauser B. (2009), Definiteness and specificity of verbal referents, in: Birzer S., Finkelstein M., Mendoza I. (eds.), Proceedings of the second international Perspectives on Slavistics conference (Regensburg 2006), Munich, p. 115–126.

Sonnenhauser B., Zangenfeind R. (2013), Towards machine translation of Russian aspect, in: Apresjan V., Iomdin B., Ageeva E. (eds.), Proceedings of the 6th International Conference on Meaning-Text Theory. Prague, pp. 192–201, available at http://meaningtext.net/mtt2013/proceedings_MTT13.pdf.

Vazov N. (1999), Context-scanning strategy in temporal reasoning, in: Modeling and using context. Lecture notes in computer science 1688, pp. 389–402.